# Analytical Gradients for Molecular-Orbital-Based Machine Learning

Sebastian J. R. Lee,[1] Tamara Husch,[1] Feizhi Ding,[2] and Thomas F. Miller III[1, 2, a)]
[1)] *Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, California 91125, United States*
[2)] *Entos, Inc., Los Angeles, CA 90027*

Molecular-orbital-based machine learning (MOB-ML) enables the prediction of accurate correlation energies at the cost of obtaining molecular orbitals. Here, we present the derivation, implementation, and numerical demonstration of MOB-ML analytical nuclear gradients which are formulated in a general Lagrangian framework to enforce orthogonality, localization, and Brillouin constraints on the molecular orbitals. The MOB-ML gradient framework is general with respect to the regression technique (e.g., Gaussian process regression or neural networks) and the MOB feature design. We show that MOB-ML gradients are highly accurate compared to other ML methods on the ISO17 data set while only being trained on energies for hundreds of molecules compared to energies and gradients for hundreds of thousands of molecules for the other ML methods. The MOB-ML gradients are also shown to yield accurate optimized structures, at a computational cost for the gradient evaluation that is comparable to Hartree-Fock theory or hybrid DFT.

## I. INTRODUCTION

Analytical nuclear gradients are the foundation of the quantum chemical elucidation of complex reaction mechanisms via molecular dynamics simulations and minimum-energy and transition-state structure optimization. However, the routine calculation of *ab initio* energies and forces with accurate wave function methods is prohibited by their steep cost, e.g., coupled-cluster singles, doubles, and perturbative triples [CCSD(T)] scales as[1] $N^7$ and full configuration interaction scales as[2] $N!$ where $N$ is a measure of system size. In recent years, machine learning has opened up a new way of mitigating the cost of quantum chemical calculations.[3–26] A particular approach that has proven to be highly data efficient and transferable across chemical space is molecular-orbital-based machine learning (MOB-ML) method.[27–29] MOB-ML relies on information of local molecular orbitals to predict the pair-wise sum of a post-Hartree–Fock correlation energy at drastically reduced cost.[27–29]

The gradient theory for MOB-ML is comparable to that of non-canonical wave function-based correlation methods, due to factors that include orbital localization and the non-variational energy expression. There exists only a handful of local wave function-based correlation methods for which this effort has been performed.[30,31] In this work, we establish a general Lagrangian framework to obtain the analytical nuclear gradients of the MOB-ML energy. The framework enforces orthogonality, localization, and Brillouin constraints on the molecular orbitals (section II B). A noteworthy aspect of this framework is that it is agnostic to the training data used for the MOB-ML model, thereby yielding accurate gradient predictions for wave function theory methods for which analytical gradients have not yet been derived or implemented. Furthermore, the computational cost of evaluating the MOB-ML energy gradient is comparable to that of a Hartree–Fock (HF) gradient or a hybrid density functional theory (DFT) gradient, such that it is orders of magnitude faster than evaluating the gradients of *ab initio* wave function theories.

We numerically validate the MOB-ML gradient theory by comparison to energy finite differences in section IV. Furthermore, we show that using only training data based on energy calculations (not gradients), MOB-ML efficiently and accurately yields gradients for diverse sets of molecules (section IV). Comparison of MOB-ML to other ML methods on the example of the ISO17 data set highlights the data efficiency and high transferability of MOB-ML for gradient predictions. We show that MOB-ML optimized structures for molecules in the ISO17 set are systematically improved with respect to the reference HF method and we compare the performance to that of a standard DFT functional.

## II. MOB-ML ANALYTICAL NUCLEAR GRADIENTS

### A. MOB-ML Energy Theory

MOB-ML relies on molecular orbital information from a HF calculation to predict a wave function correlation energy. The working equation for the

MOB-ML energy is[27–29]

$$E_{\text{MOB-ML}}[\mathbf{f}] = E_{\text{corr}}[\mathbf{f}] + E_{\text{HF}}, \tag{1}$$

where $E_{\text{HF}}$ is the HF energy, and $E_{\text{corr}}[\mathbf{f}]$ is the machine-learned correlation energy,

$$E_{\text{corr}}[\mathbf{f}] = \sum_i \epsilon_{ii}[\mathbf{f}_i] + 2\sum_{i>j} \epsilon_{ij}[\mathbf{f}_{ij}]. \tag{2}$$

The matrix of feature vectors, $\mathbf{f}$, is divided into two sub-classes. The first sub-class is made up by the diagonal components of $\mathbf{f}$, $\mathbf{f}_i$, which represent the valence-occupied orbital $i$. The second sub-class is made up by the off-diagonal components of $\mathbf{f}$, $\mathbf{f}_{ij}$, which represent the interaction between the valence-occupied orbitals $i$ and $j$. Both diagonal and off-diagonal feature vectors are composed of elements from the HF Fock matrix in the MO basis, $\mathbf{F}$, and the MO repulsion integrals, $\boldsymbol{\kappa}$, where

$$\begin{aligned}[\boldsymbol{\kappa}^{pq}]_{mn} &= (pq|mn) \\ &= \sum_{\mu\nu\kappa\sigma} C_{\mu p}C_{\nu q}C_{\kappa m}C_{\sigma n}(\mu\nu|\kappa\sigma). \end{aligned} \tag{3}$$

Here, $(\mu\nu|\kappa\sigma)$ are the four-center atomic orbital integrals with $\mu$, $\nu$, $\kappa$, and $\sigma$ representing atomic orbital indices. We restrict the MO indices of $\mathbf{F}$ and $\boldsymbol{\kappa}$ to the valence-occupied and valence-virtual MOs and we only include 2-center Coulomb- and exchange-type MO integrals, $[\boldsymbol{\kappa}^{pp}]_{qq}$ and $[\boldsymbol{\kappa}^{pq}]_{pq}$ respectively. We evaluate the feature vectors following the protocol that is specified in Ref. 32.

## B. MOB-ML Gradient Theory

### 1. Lagrangian framework

MOB-ML is a non-variational theory for which the analytical nuclear gradient theory can be derived within a Lagrangian framework,

$$\begin{aligned} \frac{\text{d}E_{\text{MOB-ML}}}{\text{d}q} &= \frac{\text{d}\mathcal{L}}{\text{d}q} = \frac{\partial\mathcal{L}}{\partial q} + \frac{\partial\mathcal{L}}{\partial\mathbf{C}}\frac{\partial\mathbf{C}}{\partial q} \\ &= \frac{\partial\mathcal{L}}{\partial q}, \end{aligned} \tag{4}$$

where $q$ refers to nuclear coordinate. The calculation of the nuclear response of the HF MOs, $\partial\mathbf{C}/\partial q$, is avoided because the Lagrangian $\mathcal{L}$ is minimized with respect to all of its variational parameters which are the MO coefficients, $\mathbf{C}$. The MOB-ML energy Lagrangian is

$$\begin{aligned} \mathcal{L}\big[\mathbf{C}, \mathbf{x}, \mathbf{z}, &\mathbf{z}^{\text{core}}, \mathbf{z}^{\text{val-occ}}, \mathbf{z}^{\text{val-vir}}, \boldsymbol{\lambda}\big] = \\ & E_{\text{MOB-ML}}[\mathbf{f}] + \sum_{pq} x_{pq}\big(\mathbf{C}^\dagger\mathbf{S}\mathbf{C} - \mathbf{I}\big)_{pq} \\ & + \sum_{ai} z_{ai}F_{ai}\Big|_{i\in\text{occ},a\in\text{vir}} + \sum_{ri} z_{ri}^{\text{core}}F_{ri} \\ & + \sum_{i>j} z_{ij}^{\text{val-occ}}r_{ij} + \sum_{ab} z_{ab}^{\text{val-vir}}r_{ab} \\ & + \sum_{wa} \lambda_{wa}P_{wa}, \end{aligned} \tag{5}$$

where $\mathbf{x}, \mathbf{z}, \mathbf{z}^{\text{core}}, \mathbf{z}^{\text{val-occ}}, \mathbf{z}^{\text{val-vir}}$, and $\boldsymbol{\lambda}$ are the Lagrange multipliers. We refer to the core MOs with column indices $r, s$, to the valence-occupied localized MOs (LMOs) with column indices $i, j, k, l$, to the valence-virtual LMOs with column indices $a, b$, and to the non-valence-virtual MOs with column indices $w, x$. The indices $m, n, p, q$ are used to index generic molecular orbitals. The first term on the right hand side (RHS) of Eq. 5 is the MOB-ML energy described by Eq. 1. The second term on the RHS constrains the HF MOs, $\mathbf{C}$, to be orthonormal, which is commonly referred to as the Pulay force.[33] The third term on the RHS is known as the Brillouin conditions, which account for the dependence of the correlation energy on the HF optimized molecular orbitals. The frozen-core conditions, $F_{ri} = 0$, account for neglecting the correlation energy contributions from the core orbitals. The localization conditions, $r_{ij} = 0$ and $r_{ab} = 0$, account for how the valence-occupied and valence-virtual MOs are localized respectively. In this work, we employ Foster-Boys localization[34] and intrinsic bond orbitals (IBO) localization,[35] but it is straightforward to generalize to other localization methods. The valence virtual conditions, $P_{wa} = 0$, reflect how the valence virtual MOs are obtained through a unitary transformation of the virtual MOs. This unitary transformation corresponds to the column space of a projection matrix formed by projecting the virtual MOs onto the IAOs. The complementary null space of this projection matrix corresponds to the non valence-virtual orbitals. This projection matrix is defined as

$$\mathbf{P} = \mathbf{C}_{\text{vir}}^{\text{IAO},\dagger}\mathbf{C}_{\text{vir}}^{\text{IAO}}, \tag{6}$$

where

$$\mathbf{C}_{\text{vir}}^{\text{IAO}} = \mathbf{X}_{\text{occ}}^{\text{IAO},\dagger}\mathbf{S}_1\mathbf{C}_{\text{vir}}, \tag{7}$$

and where $\mathbf{C}_{\text{vir}}$ is the virtual MO coefficient matrix. The matrix $\mathbf{X}^{\text{IAO}}$ transforms between the original AO and IAO basis sets and is expanded in Appendix B. All together, this yields the following an-

alytical nuclear gradient,

$$\frac{\mathrm{d}E_{\text{MOB-ML}}}{\mathrm{d}q} = E_{\text{ML}}^{(q)} + E_{\text{HF}}^{(q)} + \sum_{pq} x_{pq}\Big(\mathbf{C}^\dagger \mathbf{S}^{(q)} \mathbf{C}\Big)_{pq}$$
$$+ \sum_{ai} z_{ai} F_{ai}^{(q)}\Big|_{a\in\text{vir},i\in\text{occ}} + \sum_{ri} z_{ri}^{\text{core}} F_{ri}^{(q)}$$
$$+ \sum_{i>j} z_{ij}^{\text{val-occ}} r_{ij}^{(q)} + \sum_{ab} z_{ab}^{\text{val-vir}} r_{ab}^{(q)}$$
$$+ \sum_{wa} \lambda_{wa} P_{wa}^{(q)},$$

(8)

where the superscript $(q)$ denotes the explicit derivative of the quantity with respect to a nuclear coordinate. Eq. 8 is the general MOB-ML analytical nuclear gradient and we will now outline how to determine the Lagrange multipliers for our particular use case to arrive at a final working equation.

### 2. Minimizing with respect to MO coefficients

All of the Lagrange multipliers ($\mathbf{x}$, $\mathbf{z}$, $\mathbf{z}^{\text{core}}$, $\mathbf{z}^{\text{val-occ}}$, $\mathbf{z}^{\text{val-vir}}$ and $\boldsymbol{\lambda}$) are determined by minimizing the MOB-ML Lagrangian with respect to its variational parameters, which are the MO coefficients, $\mathbf{C}$. Differentiating the Lagrangian with respect to these parameters yields

$$\sum_\mu C_{\mu p} \frac{\partial \mathcal{L}}{\partial C_{\mu q}} = E_{pq} + 2x_{pq} + (\mathbf{D}[\mathbf{z}])_{pq}$$
$$+ (\mathbf{D}[\mathbf{z}^{\text{core}}])_{pq} + \big(\mathbf{a}[\mathbf{z}^{\text{val-occ}}]\big)_{pq}$$
$$+ \big(\mathbf{a}[\mathbf{z}^{\text{val-vir}}]\big)_{pq} + (\mathbf{D}[\boldsymbol{\lambda}])_{pq} = 0,$$

(9)

where

$$E_{pq} = \sum_\mu C_{\mu p} \frac{\partial(E_{\text{corr}}[\mathbf{f}] + E_{\text{HF}})}{\partial C_{\mu q}}$$
$$= 4F_{pq}\Big|_{q\in\text{occ}} + \big(\mathbf{F}\bar{\mathbf{D}}^{\text{F}}\big)_{pq}\Big|_{q\in\text{loc}}$$
$$+ 2\big(\mathbf{g}\big[\mathbf{C}\bar{\mathbf{D}}^{\text{F}}\mathbf{C}^\dagger\big]\big)_{pq}\Big|_{q\in\text{occ}}$$
$$+ 2\sum_m [\boldsymbol{\kappa}^{pq}]_{mm}\big(D_{qm}^{\text{J}} + D_{mq}^{\text{J}}\big)\Big|_{mq\in\text{loc}}$$
$$+ 2\sum_m [\boldsymbol{\kappa}^{pm}]_{qm}\big(D_{qm}^{\text{K}} + D_{mq}^{\text{K}}\big)\Big|_{mq\in\text{loc}}$$
$$+ \sum_{nm} R_{pm}^n\big(D_{qm}^{\text{R},n} + D_{mq}^{\text{R},n}\big)\Big|_{mq\in\text{val-occ}},$$

(10)

$$(\mathbf{D}[\mathbf{z}])_{pq} =$$
$$\sum_\mu C_{\mu p}\left(\sum_{ai} z_{ai} \frac{\partial F_{ai}}{\partial C_{\mu q}}\right)\Big|_{i\in\text{occ},a\in\text{vir}}$$
$$= (\mathbf{Fz})_{pq}\Big|_{q\in\text{occ}}$$
$$+ \big(\mathbf{Fz}^\dagger\big)_{pq}\Big|_{q\in\text{vir}} + 2(\mathbf{g}[\bar{\mathbf{z}}])_{pq}\Big|_{q\in\text{occ}},$$

(11)

$$(\mathbf{D}[\mathbf{z}^{\text{core}}])_{pq} = \sum_\mu C_{\mu p}\left(\sum_{rk} z_{rk}^{\text{core}} \frac{\partial F_{rk}}{\partial C_{\mu q}}\right)$$
$$= (\mathbf{Fz}^{\text{core}})_{pq}\Big|_{q\in\text{val-occ}}$$
$$+ \big(\mathbf{Fz}^{\text{core},\dagger}\big)_{pq}\Big|_{q\in\text{core}} + 2(\mathbf{g}[\bar{\mathbf{z}}^{\text{core}}])_{pq}\Big|_{q\in\text{occ}},$$

(12)

$$\big(\mathbf{a}[\mathbf{z}^{\text{val-occ}}]\big)_{pq} = \sum_\mu C_{\mu p}\left(\sum_{i>j} z_{ij}^{\text{loc}} \frac{\partial r_{ij}}{\partial C_{\mu q}}\right), \quad (13)$$

$$\big(\mathbf{a}[\mathbf{z}^{\text{val-vir}}]\big)_{pq} = \sum_\mu C_{\mu p}\left(\sum_{a>b} z_{ab}^{\text{vir}} \frac{\partial r_{ab}}{\partial C_{\mu q}}\right), \quad (14)$$

and

$$(\mathbf{D}[\boldsymbol{\lambda}])_{pq} = \sum_\mu C_{\mu p}\left(\sum_{wa} \lambda_{wa} \frac{\partial P_{wa}}{\partial C_{\mu q}}\right)$$
$$= (\mathbf{P}\boldsymbol{\lambda})_{pq}\Big|_{q\in\text{non-val-vir}}.$$

(15)

Eqs. 13 and 14 are expanded in Appendices A and B, respectively, $\mathbf{F}$ is the HF Fock matrix, $\mathbf{g}$ includes all of the usual HF two-electron terms, $\mathbf{R}^n$ is expanded in Appendix A, the condition $q \in \text{loc}$ restricts the sum to valence-occupied and valence-virtual MOs, $\bar{\mathbf{z}} = \mathbf{z} + \mathbf{z}^\dagger$, $\bar{\mathbf{z}}^{\text{core}} = \mathbf{z}^{\text{core}} + \mathbf{z}^{\text{core},\dagger}$, and $\bar{\mathbf{D}}^{\text{F}} = \mathbf{D}^{\text{F}} + \mathbf{D}^{\text{F},\dagger}$. The matrices $\mathbf{D}^{\text{F}}$, $\mathbf{D}^{\text{J}}$, and $\mathbf{D}^{\text{K}}$ are calculated by

$$D_{pq}^{\text{M}} = \sum_i \frac{\partial \epsilon_{ii}[\mathbf{f}_i]}{\partial \mathbf{f}_i} \frac{\partial \mathbf{f}_i}{\partial M_{pq}}\Big|_{pq\in\text{loc}}$$
$$+ 2\sum_{i>j} \frac{\partial \epsilon_{ij}[\mathbf{f}_{ij}]}{\partial \mathbf{f}_{ij}} \frac{\partial \mathbf{f}_{ij}}{\partial M_{pq}}\Big|_{pq\in\text{loc}},$$

(16)

where $M_{pq}$ refers to $F_{pq}$, $[\boldsymbol{\kappa}^{pp}]_{qq}$ and $[\boldsymbol{\kappa}^{pq}]_{pq}$, respectively. The matrix $\mathbf{D}^{\text{R,n}}$ is

$$D_{pq}^{\text{R},n} = 2\sum_{i>j} \frac{\partial \epsilon_{ij}[\mathbf{f}_{ij}]}{\partial \mathbf{f}_{ij}} \frac{\partial \mathbf{f}_{ij}}{\partial R_{pq}^n}\Big|_{pq\in\text{val-occ}}. \quad (17)$$

The partial derivatives $\frac{\partial \epsilon_{ii}[\mathbf{f}_i]}{\partial \mathbf{f}_i}$ and $\frac{\partial \epsilon_{ij}[\mathbf{f}_{ij}]}{\partial \mathbf{f}_{ij}}$ on the RHS of Eqns. 16 and 17 are the derivatives of the

machine learning prediction with respect to the feature vectors.

We emphasize that any machine learning method (e.g. Gaussian process regression, regression clustering, neural net, etc.) can be readily used in this gradient framework without modification given $\frac{\partial \epsilon_{ii}[\mathbf{f}_i]}{\partial \mathbf{f}_i}$ and $\frac{\partial \epsilon_{ij}[\mathbf{f}_{ij}]}{\partial \mathbf{f}_{ij}}$. Furthermore, we note that the following analytical nuclear gradient derivation generalizes to any type of feature-vector design and construction, so long as the feature-vector elements are obtained from $\mathbf{F}$ and $\boldsymbol{\mathcal{K}}$. The partial derivatives $\frac{\partial \mathbf{f}_i}{\partial M_{pq}}$, $\frac{\partial \mathbf{f}_{ij}}{\partial M_{pq}}$ and $\frac{\partial \mathbf{f}_{ij}}{\partial R_{pq}^n}$ are expanded in the supplementary material.

We now proceed to solve for each of the Lagrange multipliers. First, combining the stationary conditions described by Eq. 9 with the auxiliary conditions $\boldsymbol{x} = \boldsymbol{x}^\dagger$ yields the linear Z-vector equations

$$(1 - \mathcal{P}_{pq})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{pq} = 0, \tag{18}$$

where $\mathcal{P}_{pq}$ permutes the indices $p$ and $q$, which is used to solve for $\mathbf{z}$, $\mathbf{z}^{\text{core}}$, $\mathbf{z}^{\text{val-occ}}$, $\mathbf{z}^{\text{val-vir}}$ and $\boldsymbol{\lambda}$. The matrix $\boldsymbol{x}$ is then obtained as

$$x_{pq} = -\tfrac{1}{4}(1 + \mathcal{P}_{pq})\big(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}]\big)_{pq}. \tag{19}$$

The Lagrange multipliers $\mathbf{z}^{\text{val-occ}}$ are solved by considering the (valence-occupied)-(valence-occupied) part of Eq. 18, yielding

$$(1 - \mathcal{P}_{ij})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{ij} = 0. \tag{20}$$

Eq. 20 can be further simplified by showing that

$$\begin{aligned} (1 - \mathcal{P}_{ij})(\mathbf{D}[\mathbf{z}])_{ij} &= 0, \\ (1 - \mathcal{P}_{ij})(\mathbf{D}[\mathbf{z}^{\text{core}}])_{ij} &= 0, \\ (1 - \mathcal{P}_{ij})(\mathbf{a}[\mathbf{z}^{\text{val-vir}}])_{ij} &= 0, \\ (1 - \mathcal{P}_{ij})(\mathbf{D}[\boldsymbol{\lambda}])_{ij} &= 0. \end{aligned} \tag{21}$$

As a result, $\mathbf{z}^{\text{val-occ}}$ is independent of all other Lagrange multipliers, which simplifies Eq. 20 to

$$E_{ij} - E_{ji} + \sum_{k>l}(\mathcal{B}_{ijkl} - \mathcal{B}_{jikl})z_{kl}^{\text{val-occ}} = 0, \tag{22}$$

where the 4-dimensional tensor $\mathcal{B}$ is expanded in Appendix A. The set of linear system of equations defined by Eq. 22 are the Z-vector coupled perturbed localization (Z-CPL) equations which are used to solve for $\mathbf{z}^{\text{val-occ}}$. Subsequently, Eq. 13 can be used to compute $\mathbf{a}[\mathbf{z}^{\text{val-occ}}]$.

The Lagrange multipliers $\mathbf{z}^{\text{core}}$ are solved by considering the core-(valence-occupied) part of Eq. 18, yielding

$$(1 - \mathcal{P}_{ri})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{ri} = 0, \tag{23}$$

which further simplifies to

$$E_{ri} - E_{ir} + \big(\mathbf{a}[\mathbf{z}^{\text{val-occ}}]\big)_{ri} + (\mathbf{F}\mathbf{z}^{\text{core}} - \mathbf{z}^{\text{core}}\mathbf{F})_{ri} = 0. \tag{24}$$

These are the Z-vector equations used to solve for $\mathbf{z}^{\text{core}}$. Subsequently, Eq. 12 can be used to calculate $\mathbf{D}[\mathbf{z}^{\text{core}}]$.

The Lagrange multipliers $\mathbf{z}^{\text{val-vir}}$ are solved by considering the (valence-virtual)-(valence-virtual) part of Eq. 18, yielding

$$(1 - \mathcal{P}_{ab})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{ab} = 0, \tag{25}$$

which further simplifies to

$$E_{ab} - E_{ba} + \sum_{c>d}\mathcal{C}_{abcd}z_{cd}^{\text{val-vir}} = 0, \tag{26}$$

where the 4-dimensional tensor $\mathcal{C}$ is expanded in Appendix B. These are the Z-CPL equations which are used to solve for $\mathbf{z}^{\text{val-vir}}$. Subsequently, Eq. 14 can be used to compute $\mathbf{a}[\mathbf{z}^{\text{val-vir}}]$.

The Lagrange multipliers $\boldsymbol{\lambda}$ are solved by considering the (non valence-virtual)-(valence-virtual) part of Eq. 18, yielding

$$(1 - \mathcal{P}_{wa})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{wa} = 0, \tag{27}$$

which further simplifies to

$$E_{wa} - E_{aw} + \mathbf{a}[\mathbf{z}^{\text{val-vir}}]_{wa} - (\mathbf{P}\boldsymbol{\lambda})_{aw} = 0. \tag{28}$$

These are the Z-vector equations used to solve for $\boldsymbol{\lambda}$. Subsequently, Eq. 15 can be used to compute $\mathbf{D}[\boldsymbol{\lambda}]$.

Finally, the Lagrange multipliers $\mathbf{z}$ are solved by considering the virtual-occupied part of Eq. 18, yielding

$$(1 - \mathcal{P}_{ai})(\mathbf{E} + \mathbf{D}[\mathbf{z}] + \mathbf{D}[\mathbf{z}^{\text{core}}] + \mathbf{a}[\mathbf{z}^{\text{val-occ}}] \\ + \mathbf{a}[\mathbf{z}^{\text{val-vir}}] + \mathbf{D}[\boldsymbol{\lambda}])_{ai}\Big|_{a\in\text{vir},i\in\text{occ}} = 0, \tag{29}$$

which further simplifies to

$$\begin{aligned} E_{ai} - E_{ia} &+ (2\mathbf{g}[\bar{\mathbf{z}}^{\text{core}}])_{ai} \\ &+ \big(\mathbf{a}[\mathbf{z}^{\text{val-occ}}]\big)_{ai} - \big(\mathbf{a}[\mathbf{z}^{\text{val-vir}}]\big)_{ia} \\ &- (\mathbf{P}\boldsymbol{\lambda})_{ia} + (\mathbf{F}\mathbf{z} - \mathbf{z}\mathbf{F} + 2\mathbf{g}[\bar{\mathbf{z}}])_{ai} = 0. \end{aligned} \tag{30}$$

Here, the MO indices $a$ and $i$ refer to the full virtual and occupied spaces, respectively. These are the Z-vector coupled perturbed Hartree–Fock (Z-CPHF) equations. With the solutions to all Z-vector equations we can return to Eq. 19 to solve for $\boldsymbol{x}$.

### 3. Incorporating molecular-orbital localization

To provide the working expression of Eq. 8 in terms of derivative AO integrals, we must specify the molecular-orbital localization method. For this derivation, we choose the Foster-Boys and IBO localization methods to localize the valence-occupied and valence-virtual orbitals, respectively, such that

$$\frac{\partial \mathcal{L}}{\partial q} = \text{tr}\left[\mathbf{d}_a \mathbf{h}^{(q)}\right] + \text{tr}\left[\mathbf{X}_1 \mathbf{S}_1^{(q)}\right] + \text{tr}\left[\mathbf{X}_2 \mathbf{S}_2^{(q)}\right]$$
$$+ \text{tr}\left[\mathbf{X}_{12} \mathbf{S}_{12}^{(q)}\right] + \sum_n \text{tr}\left[\mathbf{W}_n (\mathbf{R}^n)^{(q)}\right] \qquad (31)$$
$$+ \tfrac{1}{2} \sum_{\mu\nu\lambda\sigma} D_{\mu\nu\kappa\sigma} (\mu\nu|\kappa\sigma)^{(q)},$$

where $\mathbf{h}$ is the standard one-electron Hamiltonian, $\mu$, $\nu$, $\kappa$ and $\sigma$ label AO basis functions in the original basis, $(\mu\nu|\kappa\sigma)$ are the two-electron repulsion integrals, $\mathbf{S}_2$ is the overlap matrix of the minimal AO basis (MINAO) used in the IBO procedure, and $\mathbf{S}_{12}$ is the overlap matrix between the original AO and MINAO basis sets. The effective one-particle density $\mathbf{d}_a$ is defined as

$$\mathbf{d}_a = \boldsymbol{\gamma} + \tfrac{1}{2}\mathbf{C}\bar{\mathbf{D}}^F\mathbf{C}^\dagger + \tfrac{1}{2}\mathbf{C}\bar{\mathbf{z}}\mathbf{C}^\dagger + \tfrac{1}{2}\mathbf{C}\bar{\mathbf{z}}^{\text{core}}\mathbf{C}^\dagger, \quad (32)$$

where $\boldsymbol{\gamma}$ is the full system HF density. The effective two-particle density $\mathbf{D}$ is defined as

$$D_{\mu\nu\kappa\sigma} = (\mathbf{d}_b)_{\mu\nu}\gamma_{\kappa\sigma} - \tfrac{1}{2}(\mathbf{d}_b)_{\mu\kappa}\gamma_{\nu\sigma}$$
$$+ 2\sum_{pq} D_{pq}^J C_{\mu p} C_{\nu p} C_{\kappa q} C_{\sigma q} \qquad (33)$$
$$+ 2\sum_{pq} D_{pq}^K C_{\mu p} C_{\kappa p} C_{\nu q} C_{\sigma q},$$

where the effective one-particle density $\mathbf{d}_b$ is defined as

$$\mathbf{d}_b = \boldsymbol{\gamma} + \mathbf{C}\bar{\mathbf{D}}^F\mathbf{C}^\dagger + \mathbf{C}\bar{\mathbf{z}}\mathbf{C}^\dagger + \mathbf{C}\bar{\mathbf{z}}^{\text{core}}\mathbf{C}^\dagger. \qquad (34)$$

The matrices $\mathbf{X}_1$, $\mathbf{X}_2$, $\mathbf{X}_{12}$, and $\mathbf{W}_n$ are defined as

$$\mathbf{X}_1 = \mathbf{C}\mathbf{x}\mathbf{C}^\dagger + \sum_{a>b} \frac{\partial r_{ab}}{\partial \mathbf{S}_1} z_{ab}^{\text{val-vir}}$$
$$+ \sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_1} \lambda_{wa},$$
$$\mathbf{X}_2 = \sum_{a>b} \frac{\partial r_{ab}}{\partial \mathbf{S}_2} z_{ab}^{\text{val-vir}} + \sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_2} \lambda_{wa}, \qquad (35)$$
$$\mathbf{X}_{12} = \sum_{a>b} \frac{\partial r_{ab}}{\partial \mathbf{S}_{12}} z_{ab}^{\text{val-vir}} + \sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_{12}} \lambda_{wa}$$

and

$$\mathbf{W}_n = \sum_{i>j} \frac{\partial r_{ij}}{\partial \mathbf{R}^n} z_{ij}^{\text{val-occ}} + \mathbf{C}\mathbf{D}^{R,n}\mathbf{C}^\dagger, \qquad (36)$$

where Eq. 35 is expanded in Appendix B and Eq. 36 is expanded in Appendix A.

### III. COMPUTATIONAL DETAILS

In this work, we perform calculations on three different data sets: (i) the thermalized water data set published in Ref. 28, (ii) a thermalized set of organic molecules featuring up to seven heavy atoms (QM7b-T),[28] and (iii) the ISO17 data set of conformers taken from molecular dynamic (MD) trajectories for constitutional isomers with the chemical formula $C_7O_2H_{10}$.[14]

All MOB-ML energy and analytical gradient are implemented in and performed with ENTOS QCORE.[36] The DF-HF calculations for the QM7b-T set,[28] and the ISO17 set,[14] are performed with a cc-pVTZ[37] basis set and a cc-pVTZ-JKFIT density fitting basis.[38] The DF-HF calculations for the water calculations are performed with a aug-cc-pVTZ[39] and a aug-cc-pVTZ-JKFIT[38] basis set. We employ a molecular orbital convergence threshold of `orbital_grad_threshold` $= 1 \times 10^{-8}$ a.u. In all MOB-ML calculations, the Foster–Boys[34] localization method is used to localize the valence-occupied MOs. The valence-virtual space is either localized with Foster–Boys localization (QM7b-T, ISO17) or the IBO localization method[35] (water). The diagonal and off-diagonal feature vectors are constructed following the procedure outlined in Ref. 32. For all Z-CPHF calculations a convergence threshold of $1 \times 10^{-8}$ a.u. is specified.

All WF calculations are performed in Molpro[40] with the frozen-core approximation, and with density fitting. All WF pair energy calculations employ the non-canonical MP2[41–44] or non-canonical coupled-cluster singles, doubles, and perturbative triples [CCSD(T)][1,45–50] correlation treatments with

6

Table I. Mean absolute error (MAE) of the MOB-ML analytical nuclear gradient with respect to the MOB-ML numerical nuclear gradient for a non-equilibrium geometry of water. The numerical nuclear gradients in were obtained with a two-step central difference formula with a step size of $5 \times 10^{-4}$ bohr. The non-equilibrium geometry of water has bond lengths of 0.986Å and 0.958Å, and a bond angle of 94.5°. All MOB-ML models are trained on data for 100 water geometries.

| Regression technique | MAE (hartree/bohr) |
|---|---|
| Linear regression | $1.45 \times 10^{-8}$ |
| Gaussian process regression | $3.75 \times 10^{-8}$ |
| Clustered Gaussian process regression | $2.28 \times 10^{-8}$ |

the cc-pVTZ, cc-pVTZ-MP2FIT,[51] aug-cc-pVTZ and aug-cc-pVTZ-MP2FIT[51] basis sets. An interface between Molpro and ENTOS QCORE is used such that WF pair energies are calculated using the DF-HF LMOs produced by ENTOS QCORE. All WF gradient calculations employ the canonical MP2 or CCSD(T) correlation treatments with the aug-cc-pVTZ, aug-cc-pVTZ-JKFIT and aug-cc-pVTZ-MP2FIT basis sets. For all Z-CPHF calculations needed for the WF gradient an iterative solver with a convergence threshold of $1 \times 10^{-9}$ a.u. is used.

The MOB-ML models for water are trained on non-canonical CCSD(T)/aug-cc-pVTZ pair correlation energies. When constructing the feature vector all non-zero elements from the Fock and $\boldsymbol{\kappa}$ matrices are used. All linear regression (LR) models are trained using Scikit-Learn.[52] All Gaussian process regression (GPR)[53] models use the Matern 5/2 kernel[53,54] and are optimized using the scaled conjugate gradient option in GPy.[55] All regression clustering models are trained following the framework outlined in Ref. 29 using a GPR within each cluster.

The MOB-ML models for the QM7b-T data set, and the ISO17 data set are trained on non-canonical MP2/cc-pVTZ pair correlation energies. Feature selection is performed using random forest regression[56] with the mean decrease of accuracy criterion, which is sometimes referred to as permutation importance.[57] All GPR models use the Matern 5/2 kernel and are optimized using the scaled conjugate gradient option in GPy.

## IV. RESULTS AND DISCUSSION

First, we compare the MOB-ML analytical gradient to the numerical gradient for an exemplary molecule to illustrate the correctness of our derivation and implementation in Table I. Table I shows that the mean absolute errors (MAE) of the analyti-

cal MOB-ML gradients of a distorted water molecule with respect to the numerical ones are on the order of $10^{-8}$ hartree/bohr for all MOB-ML models. A similar MAE is commonly found when comparing analytical and numerical gradients of pure electronic structure methods.[30,31,47,58] Additionally, Table I shows that the difference of the numerical and analytical gradient is largely independent of the regression technique (linear regression, Gaussian process regression, or a clustered Gaussian process regression) applied within the MOB-ML model. More generally, this illustrates (as also pointed out in Section II B 2) that any desired regression technique can be applied within MOB-ML without changes to the gradient framework provided that the regression prediction is differentiable with respect to the features.

As a second demonstration, we consider the thermally accessible potential energy surface of a single water molecule, following our previous work.[28] Fig. 1 shows the MAE for the energy predictions and for the associated analytical gradients we obtained with MOB-ML models trained on CCSD(T) energies performed on thermalized water geometries. As already highlighted in Ref. 29, the MAE for the energy prediction decreases steeply with the number of training geometries and we reach an MAE of $2 \times 10^{-4}$ kcal/mol when training on correlation energies of 100 training geometries. Additionally, we see that the MAE of the analytical MOB-ML gradients with respect to the analytical CCSD(T) gradients strictly decreases with an increasing amount of training data although the training data in this context are correlation energy labels and not gradients. The MAE of the MOB-ML analytical gradient is $9 \times 10^{-3}$ kcal/mol/Å when training on correlation energies for 100 water geometries. We can contextualize this result by considering that the threshold commonly used to determine if a structure optimization is converged is 0.36 kcal/mol/Å. The MAE for the gradient drops below this threshold when training on as few as three to nine water geometries. This demonstrates that MOB-ML is able to describe potential energy surfaces to a high accuracy and with a high data efficiency.

In Fig. 2, we show that this result generalizes to a diverse set of molecules. To this end, we first study the QM7b-T data set which is comprised of a thermalized set of 7211 organic molecules with 7 or fewer heavy atoms.[59] Fig. 2 shows the MAE for the MOB-ML energy prediction and for the associated analytical gradient with respect to the corresponding MP2 quantities as a function of the number of MP2 reference energy calculations.
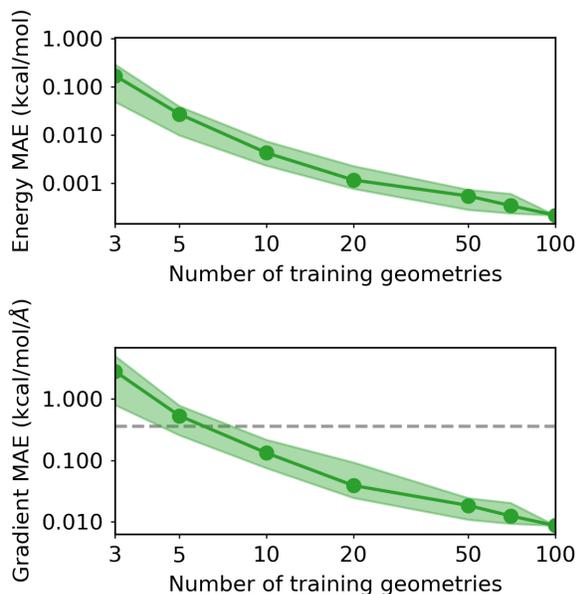
As already reported in Ref. 32, the learning curve

Figure 1. MOB-ML learning curves for CCSD(T) energies (top panel) and gradients (bottom panel) for a single water molecule. Mean absolute errors (MAE) for the predictions are reported as function of the number of water geometries used for training data; only CCSD(T) energies (not gradients) are used for the MOB-ML training data. The green circles correspond to the mean MAE obtained from 50 random samples of the training data, the green shaded area corresponds to the 90% confidence interval for the predictions and for the gradients obtained from 50 random samples. The black horizontal line at 0.3 mH/bohr in the bottom panel indicates the commonly used threshold to determine geometry optimization convergence.

for the energy decreases steeply and we obtain an MAE of 1.0 kcal/mol when training on about 70 structures. The decrease in the MAE for the energy prediction is accompanied by a decrease in the MAE for the analytical MOB-ML gradient with respect to the analytical MP2 gradient. We reach a MAE of 2.08 kcal/mol/Å when training on 220 structures.

To compare MOB-ML for gradient predictions with other machine learning methods, we now also examine the ISO17 data set.[14] The ISO17 data set consists of conformers taken from MD trajectories for constitutional isomers with the chemical formula $C_7O_2H_{10}$. Table II shows the performance of two MOB-ML models, one trained on 220 QM7b-T structures and one trained on 100 ISO17 structures, and summarizes the MAEs obtained with other ML models in the literature, i.e., SchNet,[14] FCHL,[21] PhysNet,[22] the shared-weight neural network (SWNN),[23] GM-sNN,[25] and GNNFF.[26] The MOB-ML models are the only ML models which are
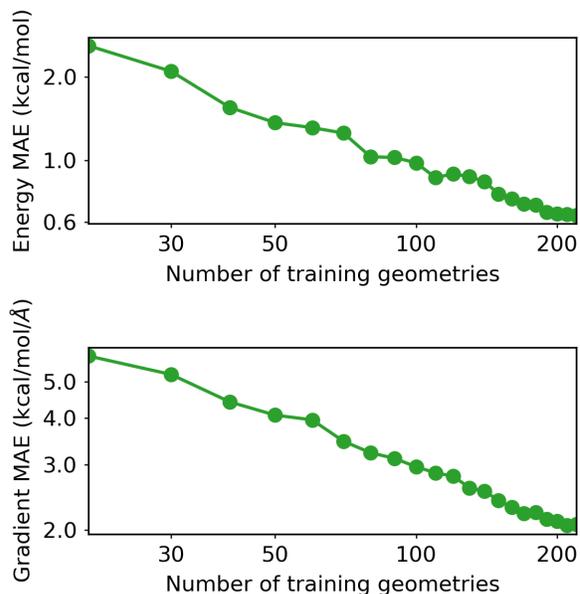


Figure 2. MOB-ML learning curves for MP2 energies (top panel) and gradients (bottom panel) for the QM7b-T data set. Mean absolute errors (MAE) for the predictions are reported as function of the number of randomly selected molecules used for training data; only MP2 energies (not gradients) are used for the MOB-ML training data.

on average chemically accurate although the MOB-ML models were only trained on energies for 100 ISO17 molecules and 220 QM7b-T molecules, respectively. The fact that our model trained on a small set of the seven-heavy atom molecules which are smaller in size than ISO17 and which are chemically more diverse (QM7b-T additionally contains the elements N, S, Cl) showcases again how transferable and data efficient MOB-ML models are. The next best model in terms of the energy MAE is GM-sNN which was trained on energies and gradients for 400k ISO17 structures and achieves an MAE of 1.97 kcal/mol. The force MAE of the MOB-ML models (1.63 and 1.64 kcal/mol/Å, respectively) is comparable to that of GM-sNN (1.66 kcal/mol/Å) while employing only 0.025% of the training data. MOB-ML is significantly more accurate in the forces than other models trained on energies alone, i.e., SchNet which obtained an MAE of 5.71 kcal/mol/Å and SWNN which obtained an MAE of 6.61 kcal/mol/Å. The only model which is more accurate in terms of the force MAE is PhysNet which is trained on energies and forces for 400k ISO17 structures. PhysNet obtains a force MAE of 1.38 kcal/mol/A. Given the demonstrated learnability of forces, it is very likely that MOB-ML could be trained to be more accurate by including more training data. Furthermore, ana-

lytical gradients have not been derived for all reference theories which considerably limits the scope of these machine learning methodologies. For example, the popular local coupled cluster methods[60–62] do not currently have derived analytical gradient theories.

Table II. Comparison of the mean absolute error for the prediction of energies and atomic forces for the unknown test set of the ISO17 data set obtained with different ML methods. The different ML methods applied different training sizes and drew on different labels to train the models on. Energy and force errors are reported in kcal/mol and kcal/mol/Å, respectively.

| Method | Training Size | Trained on energy labels | | Trained on energy+gradient labels | |
|---|---|---|---|---|---|
| | | Energy MAE | Force MAE | Energy MAE | Force MAE |
| SchNet[14] | 400,000 | 3.11 | 5.71 | 2.40 | 2.18 |
| FCHL[21] | 1,000 | — | — | 3.70 | 3.50 |
| PhysNet[22] | 400,000 | — | — | 2.94 | 1.38 |
| SWNN[23] | 400,000 | 3.72 | 6.61 | 8.57 | 6.74 |
| GM-sNN[25] | 400,000 | — | — | 1.97 | 1.66 |
| GNNFF[26] | 400,000 | — | — | — | 2.02 |
| **MOB-ML** | **100** | **0.84** | **1.64** | — | — |
| **MOB-ML** | **220**[*] | **0.76** | **1.63** | — | — |

[*]This MOB-ML model was trained on 220 randomly selected structures from the QM7b-T data set.

Despite comparing favorably to other ML methods, it remains to be shown if the MOB-ML gradients are sufficiently accurate for practical applications. Therefore, we now use the MOB-ML gradients to perform the common quantum-chemical task of optimizing molecular structures. We optimize the constitutional isomers in ISO17 with MP2 and with MOB-ML and compare the resulting structures via the root mean square deviation (RMSD) of the atoms positions in Figure 3.
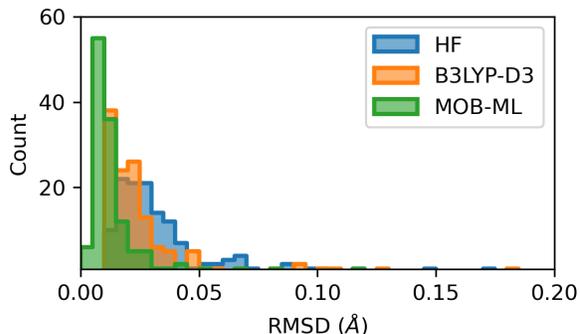


Figure 3. Histogrammed root mean square deviations (RMSD) of HF structures (blue), B3LYP-D3 structures (orange), and MOB-ML structures (green) with respect to MP2 structures for the unique isomers in the ISO17 data set. The MOB-ML was trained on 220 randomly selected QM7b-T structures.

Figure 3 shows that the MOB-ML optimized structures are very similar to the reference MP2 optimized structures with a mean RMSD of 0.01 Å. The MOB-ML optimized structures are significantly and systematically closer to the reference MP2 structures than the HF-optimized structures which exhibit an average RMSD of 0.03 Å. Moreover, the MOB-ML structures are more similar to the reference MP2 structures than those obtained from B3LYP-D3, a typical DFT exchange-correlation functional. The B3LYP-D3 structures exhibit an average RMSD of 0.03 Å with respect to the MP2 reference structures.

## V. CONCLUSIONS

In this work, we have presented the derivation and implementation of the formally complete MOB-ML analytical nuclear gradient theory within a general Lagrangian framework. We have validated our derivation and implementation by comparison of numerical and analytical gradients. The MOB-ML gradient framework can be applied in conjunction with any desired fitting technique (e.g., Gaussian process regression or neural networks) and any desired recipe for assembling the MOB-ML feature information. Furthermore, the framework for evaluating the gradient of a predicted high-accuracy wave function energy is independent of the wave function method MOB-ML was trained to predict. Hence, we can take the analytical gradient of a MOB-ML method trained to predict an arbitrary accurate wave func-

tion theory.

MOB-ML was previously shown to predict high-accuracy wave function energies at the cost of a molecular orbital evaluation. We now have shown that a MOB-ML model trained on correlation energies alone also yields highly accurate gradients for potential energy surfaces of a single molecule and for sets of diverse molecules. Specifically, we presented a MOB-ML model which obtains a force MAE of 1.64 kcal/mol/Å for the ISO17 set when only trained on reference energies for 100 molecules beating out the next best model only trained on energies in the literature, SchNet (5.71 kcal/mol/Å) which was trained on 400k molecules.[14] The transferability and data efficiency becomes even clearer when considering that we obtain an MAE of 1.63 kcal/mol/Å for the ISO17 set when training on 220 QM7b-T molecules which are smaller in size (seven versus nine heavy atoms) and which are more diverse in terms of chemical composition. The accuracy of a MOB-ML model trained on energies for 220 QM7b-T molecules for the forces is on par with some of the best ML models trained on energies and forces for hundreds of thousands of ISO17 molecules. Furthermore, we have demonstrated that a force MAE of this magnitude translates into structures which are very close to reference structures. Specifically, we obtain a mean RMSD of 0.01 Å with respect to MP2 optimized structures for the ISO17 data set which is is three times smaller than for HF or B3LYP-D3 optimized structures. Natural objectives for future work include (i) the inclusion of gradients in the training process to boost the performance in the very low data regime; (ii) the extension to an open-shell framework; (iii) the adaptation of the Lagrangian framework to derive the analytical gradients of the MOB-ML energy with respect to quantities such as electric and magnetic fields.

## SUPPLEMENTARY MATERIAL

The Supplementary Material contains the partial derivatives of the feature vector elements.

## DATA AVAILABILITY STATEMENT

The data that supports the findings of this study are available within the article and its supplementary material. The data set used in Table I and Fig. 1 is available from Ref. 59. The data set used in Fig. 2 is available from Ref. 59. The data set used in Table II and Fig. 3 is available from Ref. 14.

## Appendix A: Foster-Boys Localization

This appendix provides additional details for the Boys-related terms in Eqs. 13, 22 and 36 of the main text. The localization conditions for Foster-Boys are[31]

$$r_{ij} = \sum_n R_{ij}^n \big( R_{ii}^n - R_{jj}^n \big) = 0 \text{ for all } i > j, \quad \text{(A1)}$$

where $n$ corresponds to the x, y, z-coordinates of the position operator. The matrices $R^n$ are defined as

$$R_{ij}^n = \sum_\mu (i|n|j), \quad \text{(A2)}$$

where $|i)$ and $|j)$ are valence-occupied MOs. The orbital derivative contributions from the Foster-Boys localization conditions shown in Eq. 13 are

$$\big( \mathbf{a} \big[ \mathbf{z}^{\text{val-occ}} \big] \big)_{pq} = \sum_{i>j} \mathcal{B}_{pqij} z_{ij}^{\text{loc}} \Big|_{q \in \text{val-occ}}, \quad \text{(A3)}$$

where

$$\begin{aligned} \mathcal{B}_{pqkl} = \sum_n \Big[ & \Big( 2R_{pk}^n \delta_{kq} - 2R_{pl}^n \delta_{lq} \Big) R_{kl}^n \\ & + \Big( R_{kk}^n - R_{ll}^n \Big) \Big( R_{pl}^n \delta_{kq} + R_{pk}^n \delta_{lq} \Big) \Big]. \end{aligned} \quad \text{(A4)}$$

Next, the dipole derivative contribution from the localization conditions from the first term on the RHS of Eq. 36 is

$$\begin{aligned} (\mathbf{W}_n)_{\mu\nu} = & \frac{1}{2} \sum_{ij} C_{\mu i} C_{\nu j} z_{ij}^{\text{val-occ}} \big( R_{ii}^n - R_{jj}^n \big) \\ & + \sum_i C_{\mu i} C_{\nu i} \sum_j z_{ij}^{\text{val-occ}} R_{ij}^n. \end{aligned} \quad \text{(A5)}$$

For a full derivation of the orbital and dipole derivatives of Foster-Boys localization conditions please refer to Ref. 31.

## Appendix B: IBO Localization

This appendix provides additional details for the terms IBO-related terms in Eqs. 7, 14, 26 and 35 of the main text. The localization conditions for IBO are[63]

$$
\begin{aligned}
r_{ab} &= 4 \sum_A Q_{ab}^A \left( \left( Q_{aa}^A \right)^3 - \left( Q_{bb}^A \right)^3 \right) \\
&= 0 \text{ for all } a > b,
\end{aligned}
\tag{B1}
$$

where $A$ corresponds to an atom in the molecule. The matrices $Q^A$ are defined as

$$
Q_{ab}^A = \sum_{\mu \in A} C_{\mu a}^{\mathrm{IAO}} C_{\mu b}^{\mathrm{IAO}},
\tag{B2}
$$

where the summation over $\mu$ is restricted to basis functions at atom $A$. The matrix $\mathbf{C}^{\mathrm{IAO}}$ is the MO coefficient matrix represented in the intrinsic atomic orbital (IAO) basis which is defined as

$$
\mathbf{C}^{\mathrm{IAO}} = \mathbf{X}^{\mathrm{IAO},\dagger} \mathbf{S}_1 \mathbf{C},
\tag{B3}
$$

where $\mathbf{S}_1$ is the overlap matrix in the original atomic orbital (AO) basis and $\mathbf{C}$ is the MO coefficient matrix in the original AO basis. The matrix that transforms from the AOs to the IAOs shown in Eqs. B3 and 7 is

$$
\mathbf{X}^{\mathrm{IAO}} = \bar{\mathbf{X}}^{\mathrm{IAO}} \left( \bar{\mathbf{X}}^{\mathrm{IAO},\dagger} \mathbf{S}_1 \bar{\mathbf{X}}^{\mathrm{IAO}} \right)^{-1/2}
\tag{B4}
$$

where

$$
\bar{\mathbf{X}}^{\mathrm{IAO}} = \left( \mathbf{S}_1^{-1} + \mathbf{L}\mathbf{L}^\dagger - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\dagger \right) \mathbf{S}_{12}.
\tag{B5}
$$

The matrix $\mathbf{L}$ is the subset of the MO coefficient matrix being localized. The matrix $\tilde{\mathbf{L}}$ is

$$
\tilde{\mathbf{L}} = \bar{\mathbf{L}} \left( \bar{\mathbf{L}}^\dagger \mathbf{S}_1 \bar{\mathbf{L}} \right)^{-1/2}
\tag{B6}
$$

where

$$
\bar{\mathbf{L}} = \mathbf{S}_1^{-1} \mathbf{S}_{12} \mathbf{S}_2^{-1} \mathbf{S}_{12}^\dagger \mathbf{L}.
\tag{B7}
$$

The orbital derivative contributions from the IBO localization conditions shown in Eq. 14 corresponds to Eq. 60 in Ref. 63. The tensor $\mathcal{C}$ from Eq. 26 corresponds to Eq. 37 in Ref. 63. The overlap derivative contributions from the IBO localization conditions shown in Eq. 35 correspond to Eqs. 50 - 52 in Ref. 63.

## Appendix C: Valence Virtual Conditions

This appendix provides additional details for the terms in Eqs. 7 and 35 of the main text. In Eq. 7 the matrix $\mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO}}$ is calculated using Eqs. B4 - B7

where the matrix $\mathbf{L}$ corresponds to all occupied MOs. Next, the overlap derivative contributions from the valence virtual conditions in Eq. 35 is

$$
\begin{aligned}
\sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_1} \lambda_{wa} &= \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO}} \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO},\dagger} \mathbf{S}_1 \mathbf{C}_{\mathrm{vv}} \boldsymbol{\lambda}^\dagger \mathbf{C}_{\mathrm{nvv}}^\dagger \\
&\quad + \left( \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO}} \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO},\dagger} \mathbf{S}_1 \mathbf{C}_{\mathrm{vv}} \boldsymbol{\lambda}^\dagger \mathbf{C}_{\mathrm{nvv}}^\dagger \right)^\dagger \\
&\quad + \tfrac{1}{2} \left( \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\dagger \bar{\mathbf{X}} \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\dagger - \mathbf{S}_1^{-1} \bar{\mathbf{X}} \mathbf{S}_1^{-1} - \tilde{\mathbf{X}} \right) \\
&\quad - \bar{\mathbf{X}}^{\mathrm{IAO}} \mathbf{H} \bar{\mathbf{X}}^{\mathrm{IAO},\dagger}.
\end{aligned}
\tag{C1}
$$

The matrices $\bar{\mathbf{X}}$, $\tilde{\mathbf{X}}$, and $\mathbf{H}$ are the same as the Eqs. 56, 57 and 48, respectively, shown in Ref. 63. The evaluation of these matrices differ here by redefining the matrix $\mathbf{G}$ (Eq. 42 in Ref. 63), to be

$$
\mathbf{G} = \mathbf{S}_1 \mathbf{C}_{\mathrm{nvv}} \boldsymbol{\lambda} \mathbf{C}_{\mathrm{vv}}^\dagger \mathbf{S}_1 \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO}},
\tag{C2}
$$

the matrix $\mathbf{B}$ to be

$$
\mathbf{B} = \mathbf{X}_{\mathrm{occ}}^{\mathrm{IAO}},
\tag{C3}
$$

and the matrix $\mathbf{L}$ to span all occupied MOs. The overlap derivative contributions from the valence virtual conditions in Eq. 35 are

$$
\sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_2} \lambda_{wa} = -\tfrac{1}{2} \mathbf{S}_2^{-1} \mathbf{S}_{12}^\dagger \check{\mathbf{X}} \mathbf{S}_{12} \mathbf{S}_2^{-1},
\tag{C4}
$$

and

$$
\begin{aligned}
\sum_{wa} \frac{\partial P_{wa}}{\partial \mathbf{S}_{12}} \lambda_{wa} &= \left( \mathbf{S}_1^{-1} + \mathbf{L}\mathbf{L}^\dagger - \tilde{\mathbf{L}}\tilde{\mathbf{L}}^\dagger \right) \mathbf{V} \\
&\quad + \check{\mathbf{X}} \mathbf{S}_{12} \mathbf{S}_2^{-1}.
\end{aligned}
\tag{C5}
$$

The matrices $\check{\mathbf{X}}$, and $\mathbf{V}$ are evaluated by Eqs. 58 and 54, respectively, in Ref. 63 with the same modifications to $\mathbf{G}$, $\mathbf{B}$ and $\mathbf{L}$.

## Appendix D: Density Fitting Approximation

This appendix provides details on how the density fitting approximation can be used to approximate the four-center AO integral derivatives in Eq. 31. The AO integral derivatives are approximated by

$$
\begin{aligned}
(\mu\nu|\kappa\sigma)^{(q)} \approx (\mu\nu|\kappa\sigma)_{\mathrm{DF}}^{(q)} &= \sum_P (\mu\nu|P)^{(q)} c_{\kappa\sigma}^P \\
&\quad + \sum_P c_{\mu\nu}^P (P|\kappa\sigma)^{(q)} - \sum_{PQ} c_{\mu\nu}^P J_{PQ}^{(q)} c_{\kappa\sigma}^Q,
\end{aligned}
\tag{D1}
$$

where $P$ and $Q$ label density fitting basis functions, $(\mu\nu|P)^{(q)}$ are three-center AO integrals, and $J_{PQ}$ are two-center AO integrals. The matrix $\mathbf{c}^P$ is

$$
c_{\kappa\sigma}^P = \sum_Q \left[ \mathbf{J}^{-1} \right]_{PQ} (Q|\kappa\sigma).
\tag{D2}
$$

Substituting Eq. D1 into Eq. 31 yields

$$\sum_{\mu\nu\kappa\sigma} (\mu\nu|\kappa\sigma)_{\mathrm{DF}}^{(q)} \sum_{pq} D_{\mu\nu\kappa\sigma} =$$
$$2 \sum_{P\mu\nu} (\mu\nu|P)^{(q)} \Lambda_{\mu\nu}^P - \sum_{PQ} J_{PQ}^{(q)} \Gamma_{PQ} \tag{D3}$$

where

$$c_{pq}^P = \sum_Q \left[\mathbf{J}^{-1}\right]_{PQ} (Q|pq), \tag{D4}$$

$$\begin{aligned}
\Lambda_{\mu\nu}^P = &(\mathbf{d}_{\mathrm{b}})_{\mu\nu} \sum_{\kappa\sigma} \gamma_{\sigma\kappa} c_{\sigma\kappa}^P \\
&- \tfrac{1}{2} \sum_{\kappa\sigma} \gamma_{\sigma\mu} (\mathbf{d}_{\mathrm{b}})_{\mu\kappa} c_{\sigma\kappa}^P \\
&+ 2 \sum_q C_{\mu q} C_{\nu q} \sum_p c_{pp}^P D_{pq}^{\mathrm{J}} \\
&+ 2 \sum_q C_{\nu q} \sum_p c_{pq}^P D_{pq}^{\mathrm{K}} C_{\mu p},
\end{aligned} \tag{D5}$$

and

$$\begin{aligned}
\Gamma_{PQ} = &\sum_{\mu\nu} (\mathbf{d}_{\mathrm{b}})_{\mu\nu} c_{\mu\nu}^P \sum_{\kappa\sigma} \gamma_{\sigma\kappa} c_{\sigma\kappa}^Q \\
&- \tfrac{1}{2} \sum_{\mu\nu\kappa\sigma} (\mathbf{d}_{\mathrm{b}})_{\mu\kappa} \gamma_{\nu\sigma} c_{\mu\nu}^P c_{\kappa\sigma}^Q \\
&+ 2 \sum_{pq} c_{pp}^P c_{qq}^Q D_{pq}^{\mathrm{J}} \\
&+ 2 \sum_{pq} c_{pq}^P c_{pq}^Q D_{pq}^{\mathrm{K}}.
\end{aligned} \tag{D6}$$

## REFERENCES

[1] G. E. Scuseria and T. J. Lee, J. Chem. Phys. **93**, 5851 (1990).

[2] J. Olsen, P. Jørgensen, and J. Simons, Chem. Phys. Lett. **169**, 463 (1990).

[3] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi, Phys. Rev. Lett. **104**, 136403 (2010).

[4] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. von Lilienfeld, Phys. Rev. Lett. **108**, 058301 (2012).

[5] A. Lavecchia, Drug Discov. Today **20**, 318 (2015).

[6] E. Gawehn, J. A. Hiss, and G. Schneider, Mol. Inf. **35**, 3 (2016).

[7] P. Raccuglia, K. C. Elbert, P. D. F. Adler, C. Falk, M. B. Wenny, A. Mollo, M. Zeller, S. A. Friedler, J. Schrier, and A. J. Norquist, Nature **533**, 73 (2016).

[8] J. N. Wei, D. Duvenaud, and A. Aspuru-Guzik, ACS Cent. Sci. **2**, 725 (2016).

[9] J. S. Smith, O. Isayev, and A. E. Roitberg, Chem. Sci. **8**, 3192 (2017).

[10] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller, Sci. Adv. **3**, e1603015 (2017), publisher: American Association for the Advancement of Science Section: Research Article.

[11] E. Kim, K. Huang, S. Jegelka, and E. Olivetti, Npj Comput. Mater. **3**, 1 (2017), number: 1 Publisher: Nature Publishing Group.

[12] Z. W. Ulissi, A. J. Medford, T. Bligaard, and J. K. Nørskov, Nat. Commun. **8**, 14621 (2017).

[13] M. H. S. Segler and M. P. Waller, Chem. Eur. J. **23**, 5966 (2017).

[14] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller, "Schnet: A continuous-filter convolutional neural network for modeling quantum interactions," (2017), arXiv:1706.08566 [stat.ML].

[15] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Nature **559**, 547 (2018).

[16] N. Lubbers, J. S. Smith, and K. Barros, J. Chem. Phys. **148**, 241715 (2018).

[17] M. Popova, O. Isayev, and A. Tropsha, Sci. Adv. **4**, eaap7885 (2018).

[18] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko, Nat. Commun. **9**, 3887 (2018), number: 1 Publisher: Nature Publishing Group.

[19] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, J. Chem. Phys. **148**, 241733 (2018).

[20] J. S Smith, B. T. Nebgen, R. Zubatyuk, N. Lubbers, C. Devereux, K. Barros, S. Tretiak, O. Isayev, and A. Roitberg, "Outsmarting quantum chemistry through transfer learning," (2018), publisher: ChemRxiv.

[21] A. S. Christensen, F. A. Faber, and O. A. von Lilienfeld, J. Chem. Phys. **150**, 064105 (2019).

[22] O. T. Unke and M. Meuwly, J. Chem. Theory Comput. **15**, 3678 (2019).

[23] T. A. Profitt and J. K. Pearson, Phys. Chem. Chem. Phys. **21**, 26175 (2019).

[24] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld, J. Chem. Phys. **152**, 044107 (2020), publisher: American Institute of Physics.

[25] V. Zaverkin and J. Kästner, J. Chem. Theory Comput. (2020), 10.1021/acs.jctc.0c00347, publisher: American Chemical Society.

[26] C. W. Park, M. Kornbluth, J. Vandermause, C. Wolverton, B. Kozinsky, and J. P. Mailoa, "Accurate and scalable multi-element graph neural network force field and molecular dynamics with direct force architecture," (2020), arXiv:2007.14444 [physics.comp-ph].

[27] M. Welborn, L. Cheng, and T. F. Miller III, J. Chem. Theory Comput. **14**, 4772 (2018).

[28] L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller III, J. Chem. Phys. **150**, 131103 (2019).

[29] L. Cheng, N. B. Kovachki, M. Welborn, and T. F. Miller III, J. Chem. Theory Comput. **15**, 6668 (2019), publisher: American Chemical Society.

[30] M. Schütz, H.-J. Werner, R. Lindh, and F. R. Manby, J. Chem. Phys. **121**, 737 (2004).

[31] P. Pinski and F. Neese, J. Chem. Phys. **150**, 164102 (2019), publisher: American Institute of Physics.

[32] T. Husch, J. Sun, L. Cheng, S. J. R. Lee, and T. F. Miller III, (2020), arXiv:2010.03626.

[33] P. Pulay, Mol. Phys. **17**, 197 (1969).

[34] J. M. Foster and S. F. Boys, Rev. Mod. Phys. **32**, 300 (1960).

[35] G. Knizia, J. Chem. Theory Comput. **9**, 4834 (2013), publisher: American Chemical Society.

[36] F. R. Manby, T. F. Miller III, P. J. Bygrave, F. Ding, T. Dresselhaus, A. Batista-Romero, A. Buccheri, C. Bungey, S. J. R. Lee, R. Meli, C. Steinmann, T. Tsuchiya, M. Welborn, and T. Wiles, "entos: A quantum molecular simulation package," (2019).

[37] T. H. Dunning, J. Chem. Phys. **90**, 1007 (1989).

[38] F. Weigend, Phys. Chem. Chem. Phys. **4**, 4285 (2002).

[39] R. A. Kendall, T. H. Dunning, and R. J. Harrison, J. Chem. Phys. **96**, 6796 (1992).

[40] H.-J. Werner, P. J. Knowles, G. Knizia, F. R. Manby, M. Schütz, and others, *MOLPRO, version 2019.2, a package of ab initio programs* (Cardiff, UK, 2019).

[41] C. Møller and M. S. Plesset, Phys. Rev. **46**, 618 (1934).

[42] M. Schütz, G. Hetzer, and H.-J. Werner, J. Chem. Phys. **111**, 5691 (1999).

[43] G. Hetzer, M. Schütz, H. Stoll, and H.-J. Werner, J. Chem. Phys. **113**, 9443 (2000).

[44] H.-J. Werner, F. R. Manby, and P. J. Knowles, J. Chem. Phys. **118**, 8149 (2003).

[45] G. E. Scuseria, A. C. Scheiner, T. J. Lee, J. E. Rice, and H. F. Schaefer, J. Chem. Phys. **86**, 2881 (1987).

[46] A. C. Scheiner, G. E. Scuseria, J. E. Rice, T. J. Lee, and H. F. Schaefer, J. Chem. Phys. **87**, 5361 (1987).

[47] T. J. Lee and A. P. Rendell, J. Chem. Phys. **94**, 6229 (1991).

[48] M. Schütz, J. Chem. Phys. **113**, 9986 (2000).

[49] M. Schütz and H.-J. Werner, Chem. Phys. Lett. **318**, 370 (2000).

[50] H.-J. Werner and M. Schütz, J. Chem. Phys. **135**, 144116 (2011).

[51] F. Weigend, A. Köhn, and C. Hättig, J. Chem. Phys. **116**, 3175 (2002).

[52] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, and others, J. Mach. Learn. Res. **12**, 2825 (2011).

[53] C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning*, Adaptive computation and machine learning (MIT Press, Cambridge, Mass, 2006) oCLC: ocm61285753.

[54] M. G. Genton, J. Mach. Learn. Res. **2**, 299–312 (2002).

[55] GPy, "GPy: A gaussian process framework in python," http://github.com/SheffieldML/GPy (since 2012).

[56] L. Breiman, Mach. Learn. **45**, 5 (2001).

[57] L. Breiman, Statist. Sci. **16**, 199 (2001), publisher: Institute of Mathematical Statistics.

[58] S. J. R. Lee, F. Ding, F. R. Manby, and T. F. Miller III, J. Chem. Phys. **151**, 064112 (2019).

[59] L. Cheng, M. Welborn, A. S. Christensen, and T. F. Miller III, "Thermalized (350K) QM7b, GDB-13, water, and short alkane quantum chemistry dataset including MOB-ML features," (2019), version Number: 1.1 type: dataset.

[60] M. Schwilk, Q. Ma, C. Köppl, and H.-J. Werner, J. Chem. Theory Comput. **13**, 3650 (2017), publisher: American Chemical Society.

[61] Y. Guo, C. Riplinger, U. Becker, D. G. Liakos, Y. Minenkov, L. Cavallo, and F. Neese, J. Chem. Phys. **148**, 011101 (2018), publisher: American Institute of Physics.

[62] P. R. Nagy, G. Samu, and M. Kállay, J. Chem. Theory Comput. **14**, 4193 (2018), publisher: American Chemical Society.

[63] M. Dornbach and H.-J. Werner, Mol. Phys. **117**, 1252 (2019).